THE ELEMENTS OF SCIENTIFIC METHOD IN SOCIOLOGY

F. STUART CHAPIN
Smith College

I

At the present time sociology is largely a deductive science, if one can call an extensive and ill-defined body of knowledge a science. General principles have been deduced from the observations of a few experienced students of human nature and these principles have been elevated into theories without sufficient inductive verification. The individual phenomenon has been explained in the light of these theories. In other sciences, the progress of achievement has been in large measure due to the use of the inductive method. In the inductive method as opposed to the deductive method, the investigator passes from the examination of a considerable number of observed facts to some theory or generalization with regard to the relations existing between the observed facts. Unfortunately this has not been the procedure in sociology. There has been too much deductive philosophic generalization and far too little inductive verification.

The chief difficulty in introducing the inductive method in the science of sociology inheres in the bewildering complexity of the subject-matter with which it deals. The ultimate unit in social relations is the human individual, the most highly organized thing in organic nature. Each human being has his own individuality and differs from every other human individual. The range of characteristics possessed by the human unit is relatively wide and variation in degree is practically infinite. This diversity of individual characteristics makes it exceedingly difficult to draw valid generalizations from even the most careful observations. It would seem, therefore, that since each individual is in this sense unique and an end in himself, the only sound method of procedure is to observe each individual separately. But this is obviously impossible.

The student is consequently thrown back upon the alternative of making the most of his limited series of observations. This implies the use of some method which will enable him to determine how representative of all individuals is his limited series of observations.

The sociologist must obtain some method of recording his observations of the limited series of individuals which will reduce personal bias and individual error to a minimum. The simplest method is to count the frequency with which different degrees of a character occur. This is obviously to use the statistical method. Arthur L. Bowley says: "Statistics are numerical statements of facts in any department of inquiry, placed in relation to each other; statistical methods are devices for abbreviating and classifying the statements and making clear the relations."[1] In so far as the statistical method involves the collection of a large number of facts and the formulation of generalizations based upon the facts, it is an inductive method. The use of the statistical method necessitates the determination of a standard of measurement. The determination of standards has been of utmost importance in scientific advance. As long as standards of measurement are subjective, all is confusion. Forces are measured by their effects, not by attributing motives to them. If we try to measure some socializing force by its degree of goodness or badness, since all men differ with respect to what they consider good or bad, we shall get as many standards of measurement as we have men. Clearly we need some objective standard of measuring social phenomena. Shall we take the richest man in the world as the standard by which poverty is to be measured? Such a standard is unsatisfactory because the wide range of variation in economic status would make some people quite incapable of appreciating our standard. Evidently we need some standard of more universal acceptability.

There is an objective standard of measurement which is universally used in the statistical treatment of social phenomena—the *average*. The reasons why the *average* is such a satisfactory standard of measurement will be made clear by considering its properties. The *average* has, in general, three properties:

[1] A. L. Bowley, *An Elementary Manual of Statistics* (London, 1910), p. 1.

1. It is an objective standard. The *average* is not the result of any bias or prejudice as is often the case when standards are selected. Everyone who determines the *average* gets the same result. The *average* of a series of measurements is a quantity which is entirely separated from personal prepossession and emotion.

2. It is representative of the totality of the phenomena observed. The *average* is not obtained from any single favored measurement. It is a quantity which is relatively impartial of any single measurement although all measurements have a part in making it what it is. It is therefore representative of the group of observations.

3. It is sensitive to changes in the magnitude of the measurements which go to determine it. Many slight differences may balance a large variation.

These characteristics of the *average* are, however, relative to the kind of average used in any given case. There are three kinds of average, each quite different from the other, but all possessing in some degree the three given properties. We shall follow W. I. King's[1] enumeration of the properties of averages.

The *arithmetic average* or *mean* is the form in most common use. It may be defined as the sum or aggregate of a series of items divided by their number.[2] The items may, of course, be any kind of numerical record of observations. An important characteristic of the arithmetic average is that the sum of the differences (deviations) of all items therefrom (algebraic signs considered) equals zero. The arithmetic average has the following advantage as an objective standard of measurement for social phenomena:

1. "It may be definitely located by a simple process of addition and division and it is not necessary to arrange the data in the form of a series."

2. It gives weight to extreme deviations and it is affected by every item in the group of observations.

3. It is familiar to everyone.

On the other hand, the *arithmetic average* has certain disadvantages as compared with other forms of the average:

[1] *Elements of Statistical Method* (New York, 1912).

[2] *Ibid.*, p. 132.

1. "It cannot be accurately determined where the extremes of a series are missing."

2. "It emphasizes the extreme variations, which in most cases is undesirable." For example, the average of the series 1, 2, 3, 4, 5, 6, 7, 8, 9 is 5, and the average of the series 1, 2, 3, 4, 5, 20 is 5.83. Thus the two averages differ by less than 1 and yet the two series are essentially different, for in the second series the large item 20 quite overbalances the influence of the five smaller items whose average is 3.

3. "It is likely to fall where no data actually exist." For example, there is no number 5.83 in the second series above. It is easy to find by computation that the average number of persons in a family is 5.41 although such a number is evidently impossible.

The second form of the average is the *mode*. It is one of the most useful and important in the statistical study of social phenomena. It may be variously defined as the most frequent size of item, the position of greatest density, or the position of the maximum ordinate.[1] In Fig. 1 and Table I, the mode is the item which appears most frequently, i.e., an income of $700–$799, since in the entire series of 391 family incomes the largest number of families, 79, were found to have an income between $700 and $799 a year.[2] The *mode* has the following advantages as an objective standard for the measurement of social phenomena:

1. The mode is useful in cases in which it is desirable to eliminate the influence of extreme variations or observations which are unrepresentative. For example, in the illustration given, the income of $700–$799 is clearly more representative of the usual income in this group of observed families than the arithmetic average or mean, because in the computation of the arithmetic average the extreme items of income of $1,200 and over had undue influence in determining it.

2. "In determining the mode, it is unnecessary to know anything about the extreme items except that they are few in number." For example, as long as we know that the number of families in our

[1] A. L. Bowley, *Elements of Statistics* (London, 1901), p. 119; and King, *op. cit.*, p. 122.

[2] R. C. Chapin, *The Standard of Living among Workingmen's Families in New York City* (New York, 1910), p. 44.

observed group whose income is $1,500 and over is small, we do not need to bother about the effect they have made on the mode

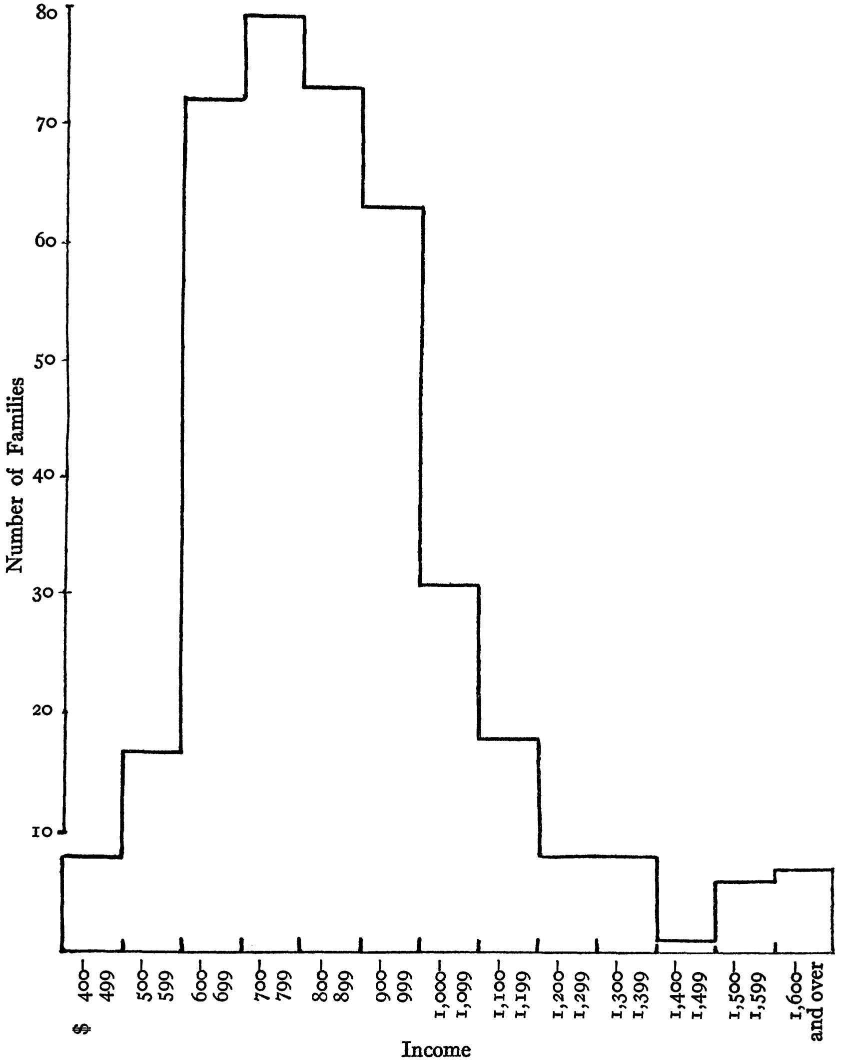

FIG. 1.—Incomes of 391 working-men's families

3. "It may be determined with considerable accuracy from well-selected sample data." For example, if Robert C. Chapin

had conducted his investigation by observing the families of Irish in one city block, instead of interviewing families scattered over various parts of the city representing the most important nationalities, his sample 391 would not have been as well selected[1] or as representative of incomes among working-men in New York City.

TABLE I*

Income	No. Families	Income	No. Families
$ 400–$ 499	8	$1,200–$1,299	8
500– 599	17	1,300– 1,399	8
600– 699	72	1,400– 1,499	1
700– 799	79	1,500– 1,599	6
800– 899	73	1,600 and over	7
900– 999	63		——
1,000– 1,099	31		391
1,100– 1,199	18		

* From R. C. Chapin, *Standard of Living among Workingmen's Families in New York City*, p. 44.

4. "The mode is a type which, to the ordinary mind, seems best to represent the group."

But the *mode* has several disadvantages which restrict its use to certain kinds of material. It is not always the best form of the average to use as a standard, because:

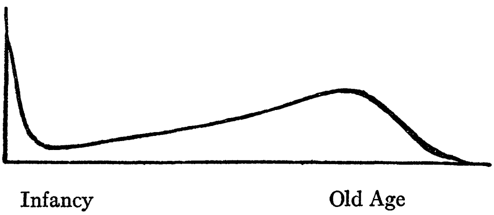

FIG. 2.—Frequency of death at different ages

1. "In many cases, no single, well-defined mode exists." Fig. 2 shows the frequency of death at different ages.[2] Here, there are two periods at which death is frequent, in early infancy and at old age.

2. "The mode is not at all useful if it is desirable to give any weight to extreme observations." In Fig. 1 the existence of 30 families with an income of $1,200 and over has no effect upon the mode.

3. "The mode may be determined by a comparatively small number of items of uniform size in a large group of varying size."

[1] Chapin, *op. cit.*, p. 28.

[2] K. Pearson, *The Chances of Death*, I, 27.

It might happen that in a community having great extremes in wealth, the modal value of possessions is $992 simply because three people were listed at that amount while the wealth of all others varied between wide limits.

A third convenient form of the average is the *median*. Bowley regards it as the most useful of averages.[1] G. U. Yule defines the median "as the middle-most or central value of the variable when the values are ranged in order of magnitude, or as the value such that greater and smaller values occur with equal frequency."[2] For

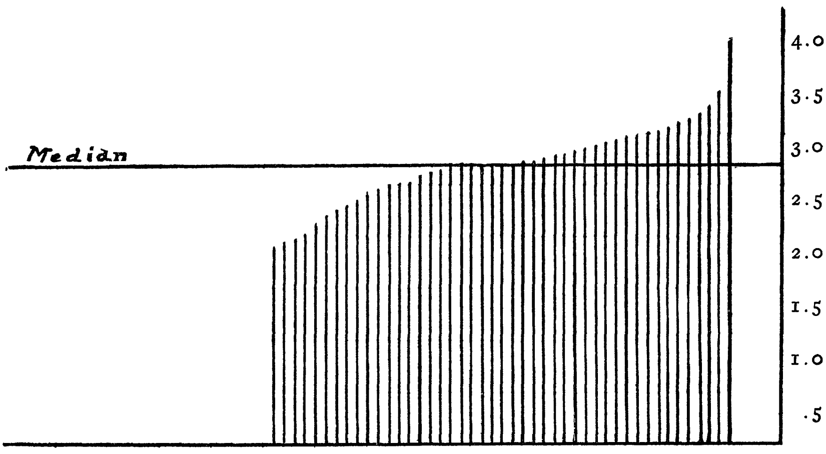

FIG. 3.—Array of breadths of 47 nuts

example, in Fig. 3, the median breadth for an observed group of 47 nuts is 2.7 cm., since this is the value half-way from either end of the ascending series of magnitudes. The median has the following advantages as an objective standard for the measurement of social phenomena:

1. "It may usually be located with greater exactitude than the mode. This is especially true in groups of observations where the mode is ill defined."

2. "It is but slightly affected by items having extreme deviations from the normal." The 6 families having an income of from

[1] Bowley, *op. cit.*, pp. 124–25.

[2] G. U. Yule, *An Introduction to the Theory of Statistics* (London, 1912), p. 116.

\$1,500 to \$1,599 do not affect the mode at all and affect the median only as much as any other single item larger than the median value would do; that is, the weight of this deviation of \$1,500 is not increased by its extraordinary size, but the item receives the same weight as any other instance and no more.

3. "Its location is never dependent upon a small number of items, as is sometimes the case with the mode."

4. "If the *number* of extreme items is known, their *size* is not required in determining the median." For example, if we know the number of persons having an income of over \$100,000, and the number of paupers, the median income could be calculated from statistics of the income of the intervening classes without considering the exact size of the income of either the very rich or the extremely poor.

5. "The median is especially useful when we are obliged to consider data the items of which are not susceptible of measurement in definite units." It is impossible to measure in concrete units the mental characteristics of a child, but it is possible to range a group of children according to their individual mentality. In such a case the arithmetic average is useless for comparative purposes, but the median can be correctly determined and its characteristics readily compared with those of other similar medians.[1]

Like the mode and the arithmetic average, the *median* has several disadvantages which must be considered in any use to which it may be put. These are:

1. In common with the mode, it is not so readily determined by a simple mathematical process as is the arithmetic average.

2. "Like the mode, it is not useful in those cases in which it is desirable to give large weight to extreme variations."

3. "Like the arithmetic average, it is often located at a point in the array at which actual items are few." For example, the median wage for the observed group might accidentally fall on the \$2.48½ per day, while perhaps only a few men actually received this amount.

[1] King, *op. cit.*, p. 131.

4. "In a discrete series in which the items are so slightly dispersed that they fall largely in the modal class, there may be so many items of the same size as the median that it becomes very indefinite."[1]

II

We have now defined our objective standard of measurement in a somewhat detailed fashion. It remains to consider the relation of this standard to the material that we wish to measure before entering upon the actual measurement. In the enumeration of the advantages and disadvantages of the different forms of the average, the terms variation and deviation were frequently used. The meaning of these terms is made clear by a consideration of our series of observations. The observed income of the 391 workingmen's families showed considerable variation, that is, all the incomes were not identical, there was a range of from $400 to $1,600 and over. The modal income, the average most representative of this group, was $700–$799, a value larger than $400 and considerably smaller than $1,600. Obviously then, the average differs from the individual items in the series from which it is obtained, and so we call all the measures which are larger or smaller than the average, variates, and the respective difference of each variate from the average, deviations. Clearly a series of observations in which there was considerable variation among the units would show large deviations. It follows from the fact that human beings are exceedingly composite units manifesting a bewildering complexity and range of characteristics, that any series of observations drawn from a group of persons will be a variable series. In working with sociological material we know that we are dealing with variables, we admit that we cannot control all the conditions in the problem, hence discrepancies between measurements are considered as due to the fact that the individuals vary from a more or less ill-defined type (the average). In experimental sciences we often assume that we are dealing with constants, hence any discrepancy between a measurement and the object is "an error of observation."

The significance of this relation between the average and the variable series of observations may be explained by reference to a

[1] *Ibid.*

recent paper by the writer, "The Variability of the Popular Vote at Presidential Elections."[1] The thesis of the paper was: increasing variability in the popular vote cast at successive presidential elections, as between states, indicates a decreasing degree of control exercised by political tradition over independence in voting. To substantiate this thesis the following method was used. A variable series was obtained by arranging in order of magnitude for any presidential election year the number of votes cast in each state. Series of this sort for Republican and Democratic presidential candidates for each year were obtained. Fourteen series, beginning with the presidential election of 1856 and concluding with that of 1908, were compared with reference to their respective variabilities around their respective medians. It was found that the fourteen Republican series and the fourteen Democratic series showed continuous and consistent increase in variability such that in the presidential elections of 1896, 1900, 1904, and 1908 the variability was over twice the variability of the year 1856. After the elimination of several considerations as to the nature of the figures and the causes at work which might lead to spurious results, the conclusion was drawn that the increasing variability in the popular vote was a real indication of increasing independence of vote and decreasing rigidity in political tradition.

The hypothesis assumed at the beginning of the investigation was that, just as increasing similarity of response to a stimulus on the part of individuals in a group indicates the slow formation of a usage or a custom of action with reference to that particular stimulus, so the increasing dissimilarity (variability) of response to a stimulus on the part of individuals in a group indicates the slow disintegration of the usage or custom. On the basis of this assumption by using a simple statistical method it seems possible to indicate the unraveling of a custom. In this particular study the stimulus was the opportunity to vote for president. It incited individuals geographically grouped by states to respond by voting for the Republican or Democratic candidate.

Instead of the popular vote for president as between states becoming standardized as time goes on, it is actually becoming diversified. We have a

[1] *American Journal of Sociology*, September, 1912.

situation in which the response of large numbers of individuals, geographically grouped, is increasingly variable with reference to a given political stimulus. If the political action of these individuals grouped by states showed increasing numerical agreement, we might say that it was due to the standardizing effect of political tradition. The fact of the matter is that the political action of these individuals grouped by states shows an increasing numerical variability and it becomes important to determine whether this increasing numerical variability is evidence of independent political action.[1]

III

We have found that the average is related to the variable series of measurements from which it is obtained in such a way that some of the items in the series are larger while some are smaller than the average. Moreover the deviations are not all of the same size. The question at once arises: Is there any law of the occurrence of these deviations? That is, do the deviations occur in a purely haphazard way with no regularity? Does each group of measurements show a series of deviations entirely different from that of preceding groups and subsequent groups? In answering this question we discover that the deviations of most measures from their averages occur with surprising regularity, that there is a definite law of their occurrence. It has been empirically demonstrated that in dealing with a large number of observations or measurements of most phenomena, when one part of the group is varying in one direction, the probabilities are that another equal part of the same group is varying in the opposite direction. Closer examination of the principle reveals the following law of occurrence of deviations of individual observations from the average of a large series of measurements:

1. Small deviations tend to occur more often than large deviations.
2. Very great deviations do not occur.
3. Deviations in one direction tend to occur as frequently as deviations in the opposite direction.

This principle will be clear by examining the distribution in Fig. 4 and Table II, representing the stature of 8,585 adult males born in the British Isles.[2] It will be seen that the average stature

[1] *American Journal of Sociology*, XVIII (1912), 223.

[2] Yule, *op. cit.*, pp. 88–89.

(modal) for the group is 67 inches. There are larger numbers of individuals with a stature of 66 and 68 inches than with a stature of 64 or 70 inches, thus fulfilling the first principle of the law. There are no individuals with statures of 24 or 120 inches, thus fulfilling the second principle. There are about as many individuals at the statures of 66 and 64 inches respectively, as at the statures of 68 and 70 inches, thus fulfilling the third principle. When these three principles are ideally realized in the occurrence of measurements, we call the resulting curve the normal curve. Our illustration is a frequency distribution which approximates somewhat closely to the ideal curve (Fig. 5). Frequency distributions of the type of Fig. 4 and more or less closely approximating the normal curve are true of large series of measurements of many human characters; for example, of the weight of men, the cephalic index, the length of infants at birth, the girth of chest of men, the strength of arm-pull of men, the body temperature at the mouth in American women, the heart-rate of American students, the reaction time of American college Freshmen, the memory span for digits in American women students, the efficiency in perception of twelve-and-one-half-year-old boys, etc.[1]

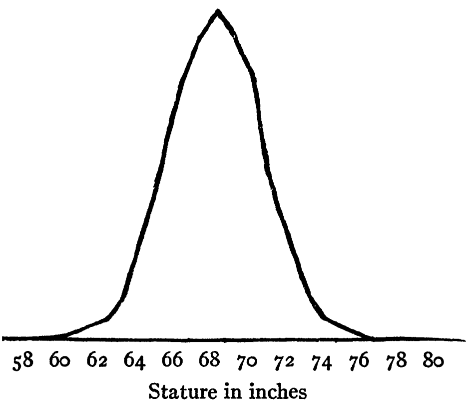

FIG. 4.—Frequency distribution of stature

The probable reason why this principle has such general applicability to measurements of organic traits is the fact that wide variation from the adapted type of inhabiting species has been strictly limited by natural selection. This can be illustrated by reference to an interesting experiment conducted by Dr. C. B. Davenport. Some 300 chickens were put in an open field; of this number 80 per cent were white or black and conspicuous, 20 per cent

[1] E. L. Thorndike, *Theory of Mental and Social Measurements* (New York, 1904), pp. 46–49.

were spotted and inconspicuous. In a short time 24 were killed by crows, but 23 of the 24 were black or white, showing that conspicuous color was a character that gave disadvantage. In due time it is probable that proportionately more of the conspicuously colored fowls would be killed, so that eventually only the spotted and inconspicuous chickens would survive. This illustrates how such a character as inconspicuousness of color favors survival, and how extreme variation (black and white) from this protective coloring (spotted) is limited by natural selection. In this way the extent to which individuals possess a trait, subject to natural selection, tends to vary within certain limits in accordance with the principles above

TABLE II*

Height in Inches	No. Men	Height in Inches	No. Men
57	2	69	1,063
58	4	70	646
59	14	71	392
60	41	72	202
61	83	73	79
62	169	74	32
63	394	75	16
64	669	76	5
65	990	77	2
66	1,223		
67	1,329		8,585
68	1,230		

* From G. U. Yule, *Introduction to Theory of Statistics*, p. 88.

outlined. A rain storm washed a large number of sparrows out of their nests. Some observers picked up the sparrows and succeeded in reviving a number of them. Both the revived and the dead sparrows were measured. It was found that the revived birds showed measurements indicating that they were more of a type than the birds killed, whose measurements were more largely unusual. In this case as in the former, Nature exterminated the extreme variates, reducing the survivors to an adapted type.

The most frequent degree of trait around which other degrees cluster, as decreasing frequencies in continuous sequence, is the type for that particular group of measurements. Thus in Fig. 4, the

typical stature is 67 inches because it is the most frequent stature found in the group of 8,585 men. Moreover, there are decreasing numbers of men as we observe successively the statures of individuals shorter or taller than the typical stature. From this point of view any individual may be regarded as a variate from some more or less well-defined type. Thus the measurement of any variable may be reduced to the determination of those elements which define the general character of the type from which it varies, or which determine the general law of distribution.[1] In the case of the statures, the frequency distribution agreed with the principles which we found to be usually true of variations from an average. Indeed, the distribution was so symmetrical that it seemed to approach some distribution of a general sort, some ideal distribution. If we could determine the characteristics of this ideal, the general law of its distribution, then we could compare our observed distribution with the ideal and determine how closely it corresponds to the ideal. The ideal might be regarded as a distribution which our observed distribution approximated but never quite equaled.

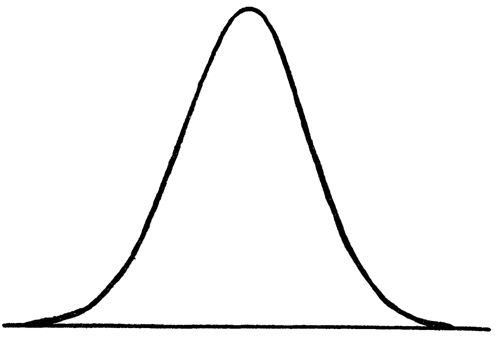

FIG. 5.—The normal curve

Mathematicians have long known that errors are distributed in accordance with three principles:

1. Small errors are more frequent than large errors.

2. Very large errors do not occur, although mathematically possible.

3. Positive errors are as frequent as negative errors.

A curve giving the distribution of errors is the normal curve of error shown in Fig. 5. The three principles comprise the law of error. We find, therefore, that there *is* an ideal distribution conforming to a law which is more or less closely approximated in

[1] F. Boas, *The Measurement of Variable Quantities* (June, 1906), p. 4.

empirical distributions when we have large numbers of items from *certain* kinds of material.

IV

In the first part of this paper we saw that because of the complexity and individuality of the units of our subject-matter it was desirable to measure each separately, but that this was impossible or at any rate impracticable; we were therefore forced to adopt the alternative of making the most of our limited series of observations and determining how representative of all individuals this limited series was. The elementary method which has been developed in the intervening paragraphs will help us toward a solution of this problem.

Clearly any limited series of observations which we have obtained in lieu of measuring all individuals, may be regarded as a sample. It becomes at once important to determine the goodness of our sample, to determine how representative it is of the larger series composed of all individuals. It is obviously impracticable for a few investigators to observe the conditions in *all* working-men's families in New York City. Resort is therefore made to the method of investigating a few representative families. Is it possible to determine how fairly the conditions in the sample 391 families represent the conditions in all working-men's families in New York City? Again, the group of 8,585 adult men from the British Isles may be regarded as a sample of all adult men in the British Isles. Is the distribution of stature characteristic of this group representative of the stature of all Englishmen? The average height of this group is 5 ft. 7 in. May we infer from this that the average height of Englishmen is 5 ft. 7 in.? In these instances as in many others it is practically impossible to measure all individuals, consequently it is of considerable importance to know whether we have a good sample or a poor one, or if we have two samples, which is the better of the two.

In the first place we might assume that in the larger group, which includes all the individuals, the measurements are distributed in frequencies which are in accordance with some general law. Thus the complete series of measurements which we cannot obtain but

which we desire to approximate as closely as possible in our limited series of observations may be regarded as an ideal. This assumption of applying the law of error is justified when we have a posteriori reasons for believing that the material we are dealing with is of a certain character. For example, we know by experiment, based upon the observation of a large number of group measurements, that biological, physiological, and anthropological measurements show a close correspondence with the law of normal distribution. It is therefore reasonable to assume, in dealing with a sample of such material, that we have a good sample when the distribution corresponds with the law of error, that we have a poor sample when the distribution fails to correspond with the law of error.

In making the assumption that the distribution of measurements in the larger group, including all individuals, is in strict correspondence with the ideal distribution of the law of error, we were justified on certain a posteriori grounds. In the absence of these grounds is it reasonable to base our method upon this assumption? The question is one of considerable importance since many measurements with which the sociologist deals are not of biological, physiological, or anthropological nature. For example, many economic phenomena show measurements which appear to obey quite different laws. In economic statistics the distribution of wealth in the population at large shows an extremely asymmetrical distribution. The percentage of population in need of relief shows a distribution less markedly asymmetrical but still failing of close correspondence with the law of error.[1]

Bowley says:

> It may appear that the cases where the agreement is close are so few as to make the whole body of theory useless; but this is an unscientific view to take. The general process of applied science is to frame hypotheses as nearly consistent with the facts as is possible without such complications as will prevent their use, and then apply to the idealized case the corrections which the actual cases necessitate. This process has led to the best results in physical science. In the problems dealt with by the law of error, it will be found that many deductions from the idealized cases hold also when applied to the only partially corresponding records of great numbers. . . . . For instance, ...

[1] Yule, *op. cit.*, pp. 92–101.

the accuracy of an average of random samples of quantities not grouped according to the curve of error varies as the square root of the number of samples taken.[1]

V

The validity of our working hypothesis as well as its relativity having been explained, it becomes important to make the practical application. When dealing with certain kinds of material we saw that, in so far as our limited series of observations corresponds to the (unlimited) ideal series, that is, in so far as the measurements in our sample were distributed in accordance with the law of error, our sample was a good one, a more or less accurate representation of the larger complete series. In many statistical problems, in most sociological problems for that matter, it is unnecessary to resort to the use of higher mathematics in order to determine the goodness of a sample. The reason for this is the fact that one cannot be sure that the statistics are accurate enough to warrant the use of refined mathematical methods. The following tests will usually be found sufficient to determine the goodness of a sample, especially in cases where there is some question of trustworthiness of the statistics:

1. The goodness of the sample depends somewhat upon its size. If our limited series of observations are few in number it is clearly improbable that the sample will be as representative of the larger series as it would be if the observations were more numerous, thus including additional numbers from the larger series and reducing the probability of any particular item being excluded. Moreover, when the sample is small we cannot in general assume that the distribution of errors is approximately normal.[2]

2. The goodness of the sample depends upon maintaining the condition that every member of the group considered has nearly the same chance of being included in the sample. That is, the sample must be *selected at random.* "The temptation is always to measure the obvious and easily accessible; but if we do this our sample is of 'the accessible,' not of the whole group. Thus the budgets of working-class expenditure, which are often published, are not

[1] Bowley, *op. cit.*, p. 298.

[2] Yule, *op. cit.*, p. 353.

typical of the working-class as a whole, but of that part of it which is intelligent enough to have some kind of record and is willing to communicate private details."[1] In other words in the returns as to family income and expenditure, the families with lower incomes are almost certain to be under-represented. Yet "it is almost impossible to say to what extent they are under-represented, or to form any estimate as to the possible error when two such samples taken by different persons at different times, or at different places, are compared."[2] If one wanted to investigate the connection between the poverty of surroundings and deformity in an individual it would be useless to go into all the poor districts of London and count the number of deformed, because there would be nothing with which to compare the result. It would not improve matters much to count all the deformed people in wealthy districts, for although we might find 5,000 in the latter case and 20,000 in the former, we "should have proved nothing until we had ascertained *how many people there were in each district.* If there were 500,000 persons residing in the wealthy districts and 2,000,000 in the poor districts, the two classes exhibit the same proportions."[3]

3. The goodness of the sample often depends upon the amount of variation among the individuals composing it. That is, the goodness of the sample depends upon the extent to which deviations from the average occur. We have assumed that deviations in the complete series obey the law of error, which implies that small deviations are more frequent than large deviations, that no very large deviations occur, and that deviations in one direction are as frequent as deviations in another direction. When the sum of the deviations (disregarding algebraic sign) of the individual items in the sample from their average is large and the total number in the sample is small, we say that the measures are considerably dispersed and do not correspond to the law of error. There is a simple statistical index that is easily computed and gives an accurate measure of the degree of dispersion. It is known as the *standard deviation* and is obtained by averaging the sum of the squares of the deviations from the average of the sample. Its formula is, $\sigma=\sqrt{\frac{\Sigma d^2}{n}}$,

[1] Bowley, *Manual of Statistics*, pp. 57–58.

[2] Yule, *op. cit.*, p. 280.

[3] W. P. and E. M. Elderton, *Primer of Statistics* (London, 1912), pp. 82–83.

where σ is the standard deviation, Σd the sum of the deviations of individual items from the average of the group, and n the total number of items in the group. When σ is large as compared with n we regard the dispersion of the sample as considerable.

By using the standard deviation we are able to compare accurately the respective dispersions of two or more samples. Other things being equal, this enables us to determine which sample is the most representative of conditions in the larger group. For example, when we have two samples to compare we compute the average of each and the respective standard deviations. The goodness of the sample is, then, a function of the number of items, and the variation among the items. More precisely, the accuracy of the average is proportional to the mean square deviation (standard deviation) and inversely proportional to the square root of the number of cases less one.[1] In symbols this is—accuracy of the average $=\frac{\sigma}{\sqrt{n-1}}$. When n is very large, the 1 may be omitted and the formula becomes $\epsilon=\frac{\sigma}{\sqrt{n}}$. The use of this formula gives us the error of the average (the goodness of the sample) and tells us how closely the average of our limited series of observations corresponds to the average of the unlimited series. In this way the measurement of our variable has been reduced to a determination of the degree in which the limited series of observations may be expected to differ from the abstract type of distribution.

Thus the precision of the average is determined as a function of the number of items and the amount of variation among them, so that one doubles the accuracy by taking four times, and trebles the accuracy by taking nine times, the number of measurements.[2] Now suppose that there is a difference between the values of the averages of the two samples, how are we to know when this difference is important? The chances that the true value lies within $\pm$ 3 times the probable error are 21 to 1.[3] Hence, whenever the difference between the means greatly exceeds these limits,

[1] Boas, *op. cit.*, p. 24.

[2] Elderton, *op. cit.*, p. 77.

[3] C. B. Davenport, *Statistical Methods* (1904), p. 14.

the discrepancy can hardly be attributed to the fluctuation of sampling and may, therefore, indicate actual differences of condition in the group from which the two samples were drawn. Thus we may be really dealing with two different groups instead of one. Consequently the way the probable error is used in practice is that, when the difference between two means exceeds three times the probable error, the difference is *significant*.[1] For example, if the difference between the mean statures of two sample groups of adult men was in excess of three times the probable error, we should think that our samples represented two different types of men, perhaps dwarfs and giants. The strict application of this method is of

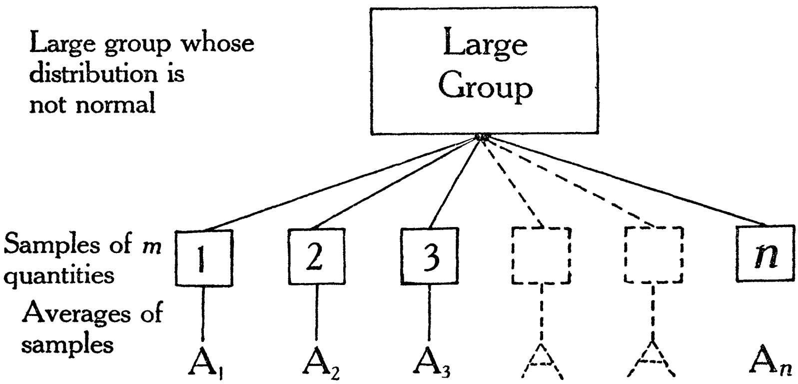

FIG. 6.—Averages of samples of a large group

course dependent upon the assumption that the groups we are dealing with are samples of quantities which conform to the law of error; the method is therefore not so well adapted to use in dealing with the samples of quantities which do not satisfy the law of error.

In dealing with quantities which do not satisfy the law of error we know that this, at least, is true, "the averages of samples of, say, m quantities, drawn at random from a large group whose distribution is not normal, will, if m is large in relation to the fluctuation of the original group, satisfy the law of error."[2] This follows from the law of probabilities which states "that a moderately large

[1] Elderton, *op. cit.*, p. 79.

[2] Bowley, *Elements*, pp. 303, 308.

number of items chosen at random from among a very large group are almost sure, on the average, to have the characteristics of the larger group."[1] Thus our method of testing samples derived from a hypothesis which is consistent with the law of error is found to apply to quantities which do not correspond so closely to the normal distribution, provided only that we fulfil the condition that our samples be large in relation to the variation in the original group. Bowley[2] has illustrated this principle by showing that even in dealing with a very unpromising case the theory is confirmed. Although the death-rates per 10,000 in London registration districts, arranged in order of magnitude, reveal a distribution which clearly does not conform to the normal curve, the averages of 18 random samples of 4 death-rates, do fit a curve of error closely.

In introducing the statistical method into the investigation of sociological phenomena we have introduced an inductive method. By means of certain assumptions based on the law of error and justified on a posteriori grounds, we have developed a means of dealing with samples of variable quantities which accurately determines, subject to certain limitations, the degree with which any sample represents the material from which it is drawn. The use of this method puts the sociologist in a position to eliminate some of the most serious difficulties arising from the complex nature of the material with which he deals. If the conditions laid down in the course of this paper are followed in applying the principles of this method to the investigation of social phenomena, it is not too much to claim that generalizations based upon the results of such investigation will be fairly comparable as regards validity and accuracy with the generalizations of applied science.

[1] King, *op. cit.*, p. 28.

[2] Bowley, *Elements*, pp. 308–15.